DOCUMENT RESUME

ED 404 367 TM 026 452

AUTHOR Haertel, Edward H.
TITLE Report on TRP Analyses of Issues Concerning Within-Age versus Cross-Age Scales for the National Assessment of Educational Progress.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE 29 Oct 91
NOTE 16p.; Some marginally legible text may not reproduce well.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Elementary Secondary Education; National Surveys; *Psychometrics; *Scaling; *Scores; *Test Results; Test Validity
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

The National Assessment Governing Board of Educational Progress has recently adopted the position that the National Assessment of Educational Progress (NAEP) should employ within-age scaling whenever feasible. The NAEP Technical Review panel (TRP) has studied the issue at some length, and reports on it in this analysis. The first section reviews the evidence concerning the tenability of the psychometric assumptions underlying cross-age (vertical) scaling, and considers whether NAEP trends or comparisons would appear materially different if within-age scaling were applied to existing NAEP data. The second section reviews the possible implications of a shift to within-age scaling for the design of the NAEP objectives frameworks and exercise pools. The third and final section relates cross-age versus within-age scaling to the substantive interpretations and policy implications supported by NAEP data. The panel concludes that in general, if one accepts the premise that cross-age scales are valid and useful, then NAEP cross-age scales are not technically flawed in any obvious ways. However, analyses suggest that cross-age scale comparisons are largely flawed and unhelpful. Overall, the report supports the recent decision of the National Assessment Governing Board to use within-age scales when feasible. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 404 367

Report on TRP Analyses of Issues Concerning Within-Age Versus Cross-Age Scales for the National Assessment of Educational Progress

Edward H. Haertel
Stanford University
October 29, 1991

Since the implementation of IRT-based proficiency scales for the National Assessment of Educational Progress (NAEP), Educational Testing Service (ETS), the current NAEP contractor, has created and reported scales spanning the three age/grade levels at which most NAEP data are collected: Ages 9, 13, and 17, or (presently) grades 4, 8, and 12. This decision has created some controversy among psychometricians, curriculum specialists, and other educational researchers, many of whom, for several reasons, believe that within-grade scales, i.e., scales defined separately for the three grade levels, would be more valid, more accurate, or less likely to be misinterpreted or misused. In response to these concerns and their own examination of the issues, the National Assessment Governing Board (NAGB) recently adopted the position that in the future, NAEP should employ within-age scaling whenever feasible.

The Technical Review Panel (TRP) examining NAEP validity issues has discussed the question of within-grade versus across-grade scales at length, and has also addressed it through a series of empirical analyses. This report on behalf of the TRP summarizes our ongoing deliberations and analyses to date.

The first section reviews evidence concerning the tenability of the psychometric assumptions underlying cross-age ("vertical") scaling, and addresses the question of whether reported NAEP trends or subgroup comparisons would appear materially different if within-age scaling were applied to existing NAEP data. The second section takes up possible implications of a shift to within-age scaling for the design of the NAEP objectives frameworks and exercise pools. The third and final section relates cross-age versus within-age scaling to the substantive interpretations and policy implications supported by NAEP data.

Validity of Current Cross-Age Scales

If an achievement test is used as intended then (apart from measurement error) a given score should represent some definite level of attainment. If the achievement level indicated by a given score depends on other attributes of the

TM026452

BEST COPY AVAILABLE

Funded by the U. S. Department of Education

examinee, then the test is biased. In particular, a score on a cross-age scale, say 300, should represent the same overall level of proficiency--and the same mix of skills--for a nine-year-old or a thirteen-year-old or a seventeen-year-old. That level of attainment by a younger versus an older child would probably be interpreted differently, of course. A proficiency level considered excellent for a fourth grader might be barely adequate for a thirteen-year-old. Nonetheless, if there is a common scale, then a given score on that scale should carry some definite implication as to what the child earning it knows or can do. In order to fulfill this basic desideratum, the NAEP cross-age scales should satisfy several requirements:

- If an exercise administered to children at more than one grade level does not function in the same way for each age group, some statistical adjustment should be made to account for its differential functioning.

- The mix of exercises functioning in a given region of the proficiency scale should be comparable across age groups with respect to the content, format, and processes assessed.

- The range of difficulty levels of exercises used at each grade level should span the range of proficiencies typically found at that grade level.

- The narrative description of each scale anchor point should apply equally to all age/grade levels found at that proficiency level. Likewise, the exercises used to illustrate each anchor point should be representative of those used at all relevant age/grade levels.

- The score distribution estimated for a given gender, racial/ethnic, or other subgroup at a given age/grade level should be virtually the same whether cross-age or within-age scaling is used.

These issues have been examined most extensively (although not exclusively) with respect to the cross-age scales used for the 1990 Mathematics assessment. ETS has made available several internal analyses, and has also provided data permitting the TRP to conduct additional analyses of its own. Evidence on each issue is reviewed in turn.

Statistical adjustment for items functioning differently at different age/grade levels. The 1990 Mathematics exercise pool consisted of 331 exercises (items), but blocks of "higher-order thinking skills" ("HOTS") and "estimation" exercises were not used in constructing the cross-age scales. Of the 275 scaled items, 88, or 32 percent, were administered at more than one age/grade level. These 88 "common" exercises (i.e., common to two or more grade levels) were examined by ETS for evidence of differential functioning across age/grade levels, using both statistical and graphical methods. Where evidence was found that different-aged children of the same overall proficiency had unequal probabilities of answering an item correctly, it was treated as if it were two separate items. Eight items (nine percent of the 88 "common" exercises) were handled in this way, and only the remaining 79 items were used to link the exercise pools used at the three age/grade levels into a common scale. No particular pattern could be detected in the eight anomalous exercises. They spanned all five of the content subareas (Numbers and Operations, Measurement, Geometry, Data Analysis, and Algebra) and all three of the process categories (Conceptual Understanding, Procedural Knowledge, and Problem Solving), and included both multiple-choice and constructed response exercises. Their difficulties ranged from the 200 through the 300 levels.

In conclusion, a significant proportion of the NAEP mathematics exercises used at more than one age/grade level functioned differently for children at different levels. Appropriate statistical adjustments were made by estimating separate sets of parameters for these items for different age cohorts. It is important to point out that neither ETS staff nor the members of the TRP could find any systematic explanation for the anomalous functioning of these exercises. Clearly, if cross-age scaling is used in the future, the same or comparable item-level examination of fit will be called for.

Comparable exercise mix. In discussing the application of item response theoretic (IRT) models to the NAEP exercise pools, Robert J. Mislevy has observed that these models are never strictly correct. He recommends that the (unidimensional) IRT model be viewed as no more than a useful and efficient tool for data reduction and description. Under his view, the proficiency scale represents an average over a sort of "market basket" of knowledge and skills, analogous to the market basket of goods used to define the Consumer Price Index. Just as the prices of different commodities may rise or fall at different rates, so patterns of performance for

BEST COPY AVAILABLE

different content areas or processes may vary. The meaning of the scale score depends on the mix of knowledge and skills included. NAEP exercises are not really unidimensional, even within content subareas summarized using unidimensional IRT models. If scores of 250, say, earned by nine-year-olds and by 13-year-olds are to mean the same thing, then it is imperative that they be based on mixes of exercises that are closely similar with respect to content, process, and format. Note that under NAEP's Balanced Incomplete Block (BIB) matrix sampling plan, even two children in the same age cohort are likely to respond to different exercises. For most interpretations of the proficiency scales, however, the comparability of the total pools at different age/grade levels is more important than the comparability at the level of exercise blocks or booklets.

The mix of content and process categories represented in different regions of a vertically equated scale may differ considerably, which is one reason many psychometricians find vertically equated scales suspect. Indeed, the framework for the 1990 NAEP mathematics assessment specifies that the allocation of exercises to different content subareas should vary across age/grade levels. The validity requirement examined here, however, is more focused. The TRP carried out detailed analyses of the NAEP exercise pool to determine whether exercises functioning within a narrow region of the proficiency scale were comparable across age/grade levels.

Of the 275 scaled exercises in the 1990 mathematics exercise pool, 135 were classified as "anchor" items, meaning that they sharply distinguished between successive scale points, and were central to the definition of the anchor points at levels 200 (27 items), 250 (30 items), 300 (48 items) or 350 (30 items). An additional 53 items that came close to meeting the anchoring criteria, referred to here as "almost-anchor" items, could also be classified into one of these four scale regions (8, 16, 16, and 13 items at levels 200, 250, 300, and 350, respectively). Within each scale region, anchor and almost-anchor items used at successive age/grade levels were compared as to content subarea, process category, and format (multiple-choice versus open-ended). Systematic differences with respect to any of these characteristics would have implied differences in the meaning of scale scores in that region as a function of age/grade cohort.

The results of these analyses indicated acceptable homogeneity across age/grade cohorts with respect to all attributes, at all scale levels. None of the chi square tests comparing patterns of exercise classifications across

cohorts reached statistical significance, and the occasional marginally significant anomalies showed no systematic pattern.

These results may be no more than a fortuitous consequence of other constraints on the composition of the exercise pool. As already noted, 32 percent of the scaled exercises were used at two or even all three age/grade levels, assuring some degree of commonality even if all of the exercises specific to single age/grade cohorts diverged. In addition, all content subareas had to be well represented among the common exercises and at each age/grade level to permit both cross-age IRT scaling and accurate performance estimation for each separate subarea. Finally, the objectives framework required that at each of the three age/grade levels, 40 percent of the exercises were to assess conceptual understanding, 30 percent procedural knowledge, and 30 percent problem solving, further assuring coherence. Even with these constraints, of course, it was logically possible that exercises at within a given region of the proficiency scale would differ from one age/grade cohort to another, but it was unlikely. Nonetheless, if cross-age scales are to be constructed in the future, similar tabulations should be made to assure that the mixes of processes, skills, and exercise formats in each scale region are similar across age/grade groups.

Span of difficulty levels. A cross-age scale has the potential to mask "floor" and "ceiling" effects, because the difficulty range of the entire reporting scale may extend above or below that of a single age/grade cohort's exercise pool. Again using the 1990 mathematics assessment as an example, the fourth grade exercise pool included only 9 exercises at level 300, and none at level 350. If significant numbers of fourth grade students had performed at these levels, their score distribution would have been poorly determined in this region of the scale. Moreover, interpretive statements that some finite proportion of fourth graders knew 350-level content or could do 350-level work would have been problematical, because no fourth graders were in fact asked to attempt such work. This argument may appear circular, but it is not. The levels at which exercises are located refer to the scale regions in which they function best, but fourth graders doing very well on the (easier) exercises administered to them could in principle earn scores in the higher region.

BEST COPY AVAILABLE

The State of Mathematics Achievement report (p. 55) indicates that only 11 percent of fourth grade students were at or above level 250, and 0 percent were at or above level 300, so in fact, no such "ceiling" problem arose. Likewise, even though twelfth graders received only 4 exercises functioning at level 200, 100 percent of them were found to be above that level (although 9 percent were below level 250).

In summary, for each of the three age/grade cohorts, the TRP comparisons of exercise difficulty ranges to examinee proficiency ranges indicated sufficient floor and ceiling to permit accurate ability estimation. Clearly, the true proportion of fourth graders at level 300 or 350 is greater than zero. Some tiny fraction of exceptionally gifted children who have received special educational advantages can probably perform at that level. Similarly, there are surely some few in-school 17-year-olds still struggling with 200-level work in mathematics. But a national assessment must be concerned with estimating achievement for the vast majority of students, not these tiny minorities. Greater breadth in exercise difficulty could only be attained by lengthening the assessment or by reducing the number of exercises and thereby the accuracy of the assessment in the difficulty ranges now assessed.

Anchor point descriptions. If a given scale score is to carry the same implications as to what the child earning it knows or can do, then it is essential that the narrative descriptions of scale anchor points apply equally to children from different age/grade cohorts who perform at that scale level. Accordingly, the TRP analyses included tabulations of anchor exercises for each of scale points 200, 250, 300, and 350 according to the age/grade levels at which they were administered. Similar tabulations were made of just those anchor exercises displayed in NAEP reports to communicate the meaning of the four anchor proficiency levels. These tabulations confirmed that anchor exercises for all relevant proficiency levels were well represented in each of the three exercise pools. Moreover, those anchor exercises displayed to express the meaning of scale levels appeared representative of exercises administered to all three age/grade cohorts. (Displayed exercises included from 39 percent to 46 percent of the anchor exercises administered to each age/grade cohort. Comparisons of displayed and nondisplayed anchor exercises also confirmed that the exercises displayed were representative with respect to content subarea, process, and format. These tabulations were

carried out for each anchor point separately as well as across anchor points.)

Of course, the actual narrative descriptions of anchor points are one step further removed from the exercise pools themselves. It is likely that some specific phrases in anchor point descriptions were derived from examination of just one or two exercises that were not used with all relevant age/grade cohorts. No evidence was found, however, suggesting any strong or systematic lack of applicability of any anchor point description to age/grade cohorts performing in the region of that anchor point.

Empirical comparisons of NAEP cross-age versus within-age scaling. Although ETS has focused primarily on cross-age proficiency scales for reporting, their analyses have occasionally required within-age scaling, which has afforded the opportunity for direct comparisons of cross-age and within-age scaling results. The discussion in this section is based on reports by ETS personnel of such comparisons.

In the 1986 mathematics assessment, in addition to the main ("cross-sectional") sample, smaller "bridge" samples were tested at concurrently, following slightly different procedures, in order to maintain accurate trend lines. Within-age scales were created for the bridge samples, and these within-age scales were then equated to the (across-age) scale estimated for the main sample. The equating procedure used assured that cross-age and within-age scales would yield identical estimates of the mean and standard deviation for an entire grade cohort, but means for subgroups were not so constrained. Thus, comparing within-grade and across-grade means for males and for females; for Anglo, Black, Hispanic, and Asian students, etc., provides an indirect test of the psychometric assumptions of the cross-age scale. Johnson, Yamamoto, and Mazzeo (no date) report such comparisons for gender and racial/ethnic categories for nine-year-olds, which show only very small differences between the results of the two scaling procedures for any of these subgroups. They assert that "similar results were obtained for the other age-levels" (p. 3).

A second opportunity for cross-age versus within-age scaling comparisons arose in the course of analyzing the 1990 reading trend data. In order to link the 1990 data to the original 1984 reading scale, it proved necessary to re-calibrate the 1988 exercise pool. Donoghue (no date) reports comparisons of 1988 reading performance estimated using item parameters obtained in this (within-grade) recalibration versus item parameters from the original (vertically equated)

BEST COPY AVAILABLE

1984 scale. Means and standard deviations are compared for the population as a whole as well as males and females; Whites, Blacks, and Hispanics; and examinees in the Northeastern, Southeastern, Central, and Western regions of the United States, for ages 9, 13, and 17. These comparisons again show only small differences between the results of the two scaling procedures.

A third opportunity to compare cross-age and within-age scales arose in connection with the 1990 Trial State Assessment (TSA) in eighth grade mathematics. From the National sample, the subset was formed of public-school eighth grade students who lived in those states participating in the TSA and who were tested at about the same time as the TSA data collection. With appropriate weighting, this "State Aggregate Comparison" (SAC) sample represented the same population as the combined TSA samples, excluding Guam and the Virgin Islands. Population and subpopulation score distributions were estimated from the SAC sample based on the National-NAEP cross-age scaling, and for the pooled TSA sample according to an independent within-grade scaling. Following a global common-population equating, comparisons of means for gender, racial/ethnic, and paternal education categories revealed no statistically significant differences between the results of the two procedures.

In a sense, all of these comparisons are fairly weak. If substantial differences had been found, they would not necessarily be attributable to within-age versus across-age scaling per se, and the fact that no significant differences were found does not establish the validity of the vertically equated scales. The 1986 mathematics comparisons involved tape-paced versus self-paced exercises, and the 1990 SAC-TSA comparisons confound scaling procedure with various small differences in the administration procedures and possibly the motivational context of the respective assessments. The 1988 reading data may provide the most sensitive test, because exactly the same data are compared, analyzed according to two different procedures. Here, however, the comparison less than optimal because the vertical scaling was originally done using data collected in 1984, and because the within-grade calibration was done scaling together data from 1984, 1988, and 1990.

Even though most subgroup differences were small, there appeared to be a tendency, across these separate examinations, toward somewhat larger differences between methods for Hispanics than for any other group. In the reading trend data, for example, the within-grade scales yielded Hispanic means 1.58 points higher and 2.24 points lower than the across-grade scale at ages 13 and 17,

respectively. The standard errors of these differences are difficult to calculate, because the estimates, based on the same data, are obviously highly correlated. In the 1986 mathematics comparisons at age 9, the largest differences for any groups reported were 1.9 standard errors for Blacks and 1.7 standard errors for Hispanics. The SAC-TSA comparison yielded a 4-point discrepancy for Hispanics, larger than for any other group except the very small American Indian sample, although still not statistically significant (t=1.75). It is possible that these larger differences for Hispanics reflect some interaction of age and language proficiency. The question may be asked for any group whether older versus younger children who earn the same overall score do so in the same way, whether they show the same pattern of performance. This question may be more important for nonnative English speakers, for whom an achievement test may measure language proficiency and/or specialized language skills along with whatever it is intended to assess.

Summary of validity for current cross-age scales. As noted above, some critics of the NAEP cross-age scales have asked whether in principle it is possible to create valid and meaningful scales spanning such a broad age range, and others have questioned the utility of such scales, regardless of their technical adequacy. The analyses reported above do not resolve these questions. If one accepts the reasonableness of cross-age scales, then the ETS implementation of cross-age scaling procedures for the 1990 mathematics assessment appears satisfactory, as do the results of selected examinations of 1986 and 1988 NAEP mathematics and reading data. The difficulty ranges of separate exercise pools cover the relevant populations; common exercises are reasonably representative of the total exercise pool with respect to content, process, and format; common exercises that for whatever reason function differently for different ages are detected and scaled separately within age; anchor and almost-anchor items are well represented in the separate exercise pools; those anchor exercises displayed in NAEP reports do in fact characterize the exercise pools at all age/grade levels to which they pertain. Where incidental comparisons have been made of the results of cross-age and within-age scaling procedures, they have shown a high degree of comparability, although results for Hispanic subpopulations tend to show a persistent sensitivity to statistical method.

NAEP exercise development is subject to many constraints, which probably helped to assure the homogeneity of separate cohorts' exercise pools at specific proficiency levels, as

BEST COPY AVAILABLE

well as the representativeness of anchor items and displayed items. Nonetheless, it is possible that some of the benign results reported here were somewhat fortuitous. Some deliberate attention to these issues for any future cross-age scales will help to assure their continued quality.

Implications of Cross-Age Versus Within-Age Scaling for Objectives Frameworks and Exercise Pools

In the 1990 mathematics assessment, 32 percent of the 275 scaled exercises were administered to groups of children at more than one age/grade level. If estimation and HOTS exercises are also considered, there were 133 common exercises out of a total pool of 331 exercises, just over 40 percent. If the same numbers of exercises had been administered at each age/grade level with no overlap, the total pool would have grown from 331 to 537 exercises, an increase of over 60 percent. It is reasonable to expect that the most difficult mathematics exercises given to fourth graders might be the same as the easiest exercises for eighth graders, or that the most difficult eighth grade items would be appropriate to include in the twelfth grade exercise pool, but 40 percent overlap seems surprising. Moreover 22 of the 275 scaled exercises (8 percent) were used with both nine-year-olds and 17-year-olds. (All but one of these 22 exercises were also used with 13-year-olds.)

The ETS exercise development process does not formally or explicitly include the goal of creating exercises suitable for use at more than one age/grade level, but a substantial number of common exercises are obviously necessary if cross-grade scaling is to be used. Moreover, substantial savings must accrue from such a dramatic reduction in the total number of exercises required, relative to the number that might be created if the three exercise pools were more-or-less independent.

A change from cross-age to within-age scaling could have significant implications for the design of objectives frameworks and creation of exercise pools, but it is difficult to say just how the frameworks and items would change. Frameworks and exercises would not necessarily change at all, but a shift to within-age scaling would remove a major constraint that at least implicitly shapes the NAEP design, providing substantially greater flexibility. How to exploit that flexibility would be a policy decision, to be informed by broader assessment purposes. That being said, the following changes might occur:

- Substantially fewer exercises would be common to more than one age/grade level, and substantially more exercises would be specific to each age/grade level.

- A broader range of age-appropriate knowledge and skills could be included.

- Different content subareas could be specified at each age/grade level according to the curriculum distinctions most meaningful at that level.

- NAEP might be designed to show greater "discriminant validity", i.e., greater sensitivity to distinct performance patterns for different content subareas.

Substantive Interpretations of Across- Versus Within-Grade Scales

The validity of NAEP uses and interpretations depends on far more than the technical adequacy of NAEP's scaling procedures. A major part of the debate between proponents of within-age versus cross-age scales has centered on the forms of interpretations enabled by the two types of scales. In general, any interpretation that can be made using within-grade scales can also be made using across-grade scales, but not conversely. Vertical equating invites some forms of interpretation that are difficult or impossible to reach using within-grade scales. Discussions of the cross-age scaling question, including deliberations within the TRP, has centered on whether these additional forms of interpretation are technically defensible and if so, whether they are sufficiently useful to justify the technical complexities, exercise pool constraints, and risks of other less appropriate interpretations engendered by cross-age scaling. Some of these forms of interpretation are as follows. "Quantitative" interpretations are those involving a direct quantitative statement embodying comparisons that depend on cross-age scaling. "Curricular" interpretations are less direct, and tend to involve textual, as opposed to purely numberical, interpretations of NAEP performance.

Quantitative Interpretations:

- Interpretation of within-grade contrasts or trends in terms of "grade equivalents"

- Comparisons of rates of growth for different age/grade levels

BEST COPY AVAILABLE

- Comparisons of the size of gaps among subgroups at different age/grade levels

- Comparisons of high-performing subgroups at one age/grade level with low-performing subgroups at another

Curricular Interpretations:

- Conception of curriculum as a single, linear progression from low-level "tool skills" through higher-level "applications"

- Conception of scale anchor points in terms of contents and processes common to broad ranges of grade levels

Interpretations in terms of "grade equivalents." This interpretation, along with several others, depends critically on the linearity of the cross-age scale. Perhaps the best illustration was given by the 1986 Reading anomaly, wherein a change of about 3 percent in the probability of a 17-year-old's answering a reading exercise correctly was translated to a drop of "a full grade level" in 17-year-old reading proficiency between 1984 and 1986. This figure was reached by taking the difference in overall mean scale scores for 13-year-olds and 17-year-olds, treating this as the gain to be expected over four years, and dividing by four to define expected annual growth. The "grade level" metric made a very small absolute change in performance appear much more substantial. Because 13-year-olds and 17-year-olds are typically tested on different content, very strong assumptions are entailed in expressing the difference between 1984 17-year-olds' performance (on grade 12 reading) and 1986 17-year-olds' performance (on grade 12 reading) in terms of the difference in scale scores corresponding to 13-year-old performance (on grade 8 reading) and 17-year-olds' performance (on grade 12 reading).

Comparisons of growth rates. On page 55 of The State of Mathematics Achievement the statement appears,

> As would be expected, twelfth graders had higher average proficiency than did eighth graders, who in turn performed better than fourth graders. Eighth graders performed, on average, 50 points higher on the scale than did fourth graders. The twelfth graders, however, on average, performed only 30 points higher on the scale than did the eighth graders.

This statement at least implicitly suggests that growth in mathematics proficiency is more rapid between grade 4 and grade 8 than between grade 8 and grade 12. No further interpretation is offered, but the reader's attention is directed to the scale point descriptions, which characterize performance at levels 200, 250, 300, and 350. Inspection of the scale point descriptions highlights the fragility of any "equal interval" interpretation for the NAEP proficiency scale. In what sense is the distance from the 200 description to the 250 description the same as the distance from the 250 description to the 300 description, for example? In fact, it is very difficult to say anything useful about the fact that eighth graders outperform fourth graders by more points than twelfth graders outperform eighth graders.

Performance gaps growing larger or smaller. Closely related to the foregoing interpretation is the comparison of subgroup differences at one grade level versus another. Examples might include interpretations of whether the gender gap in mathematics achievement is larger or smaller at grade 12 than at grade 8 or whether the effect of private versus public schooling grows or diminishes over time. As with other interpretations that depend on the linearity of the cross-age scales, these interpretations are problematical. There are, however, metric-free methods of making such distributional comparisons, which are equally straightforward to implement and interpret using within-grade as using across-grade scales. One could, for example, determine the quantile of the mathematics achievement distribution for males that corresponded to the median of the distribution for females and conversely, at each of grades 8 and 12, and compare the results to assess movement toward or away from parity.

Comparing high performers at one grade level with low performers at another. In some content areas, scale score distributions for different age/grade levels may overlap sufficiently that mean scores for different groups at different grades are comparable. For example, a recent Science Report Card pointed out the similarity in means for Black 17-year-olds versus White 13-year-olds. As discussed above, such a comparison is only valid if the patterns of knowledge and skill for these two groups are similar. But standard Differential Item Function (DIF) analyses are unlikely to include direct comparisons of different racial/ethnic subgroups at different grade levels. Significant differences in the probabilities of correct

BEST COPY AVAILABLE

responses to particular exercises might easily go undetected. Moreover, differences in general test taking skill, motivation, and general knowledge as well as science knowledge may have interacted in complex ways to yield superficially comparable performance levels. A final objection to this form of interpretation is that it is useless for informing curriculum, instruction, or educational policy. Reliance on within-grade scales would direct attention to more productive and meaningful comparisons among groups of children of the same age.

Linear conception of curriculum. Cross-age scaling may encourage a view of curriculum and learning in terms of progress along simple, unidimensional continuua spanning (at least) grades 4 through 12. Such a view tends to support the idea that advanced, higher-order skills must be reserved for the later years of schooling, and children during their earlier years need to concentrate on largely meaningless, decontextualized "tool" skills in preparation for that later application. An alternative conception (and scaling) of curriculum within grade levels can direct attention to higher-level application and problem solving for younger children as well as older, and can provide assessment information more in keeping with current reform initiatives in various curriculum areas.

Anchor point descriptions. The use of anchor point descriptions intended to apply equally to fourth, eighth, and twelfth grade students at a given proficiency level is problematical for the same reason as the use of cross-age proficiency scales. The location of different skills and abilities is largely determined by the conventions of curriculum organization, so that children at a given grade level are necessarily confined to a relatively narrow scale score region, and consequently, a very limited number of anchor point descriptions. Within-grade scales with separate anchor point descriptions for fourth, eighth, and twelfth grade students would depict more clearly the range of achievement levels and variety of attainment patterns characterizing different subgroups within each grade level, and would not divert attention to largely meaningless comparisons between the knowledge and skills of children four or eight years apart in age.

Summary

This report summarizes analyses and discussions by members of the TRP, as well as reports of relevant analyses conducted by ETS staff. It demonstrates several types of simple analyses to assure the validity of cross-age scales, which might be incorporated in future studies by the NAEP contractor. It is concluded that in general, if one accepts the premise that cross-age scales are valid and useful, then NAEP's cross-age scales are not technically flawed in any obvious ways. The report then goes on to consider the costs of cross-age scaling in terms of constraints on objectives frameworks and exercise pools, and concludes that substantially greater flexibility would result from within-age scaling. Recommendations as to how that flexibility should be exploited are beyond the scope of this paper. Finally, illustrative interpretations of cross-age scales are critiqued, and the substantive basis of such scales is called into question. Cross-age scale comparisons are found to be largely flawed and unhelpful. Overall, the report fully supports NAGB's recent decision that within-age scales should be used whenever feasible.

BEST COPY AVAILABLE